Published in final edited form as:

Phys Med Biol. 2013 November 7; 58(21): 7481–7496. doi:10.1088/0031-9155/58/21/7481.

Automatic 3D Ultrasound Calibration for Image Guided Therapy Using Intramodality Image Registration

Jeffrey Schlosser^{a,b,#}, Can Kirmizibayrak^{c,#}, Vijay Shandasani^d, Steve Metz^d, and Dimitre Hristov^{c,*}

^aDepartment of Mechanical Engineering, Stanford University, Stanford CA 94305

^bDepartment of Bioengineering, Stanford University, Stanford CA 94305

^cDepartment of Radiation Oncology, Stanford University, Stanford CA 94305

^dPhilips Ultrasound, Bothell, WA 98021

Abstract

Many real time ultrasound (US) guided therapies can benefit from management of motion-induced anatomical changes with respect to a previously acquired computerized anatomy model. Spatial calibration is a prerequisite to transforming US image information to the reference frame of the anatomy model. We present a new method for calibrating 3D US volumes using intramodality image registration, derived from the “hand eye” calibration technique. The method is fully automated by implementing data rejection based on sensor displacements, automatic registration over overlapping image regions, and a self-consistency error metric evaluated continuously during calibration. We also present a novel method for validating US calibrations based on measurement of physical phantom displacements within US images. Both calibration and validation can be performed on arbitrary phantoms. Results indicate that normalized mutual information and localized cross correlation produce the most accurate 3D US registrations for calibration. Volumetric image alignment is more accurate and reproducible than point selection for validating the calibrations, yielding <1.5 mm root mean square error, a significant improvement relative to previously reported hand eye US calibration results. Comparison of two different phantoms for calibration and for validation revealed significant differences for validation ($p=0.003$) but not for calibration ($p=0.795$).

Keywords

ultrasound (87.63.dh); ultrasound calibration; image guidance; image registration (87.57.nj); radiation therapy (87.53.Kn)

II. Introduction

Physical motion of an ultrasound (US) transducer during imaging yields apparent displacements of stationary structures within the US image. For real-time image guidance

*Corresponding author dimitre.hristov@stanford.edu.

#These authors contributed equally to this work

Note: Addresses above represent author employment when the article was prepared. Jeffrey and Can are currently employed elsewhere:

Jeffrey Schlosser is with SoniTrack Systems, Inc., Mountain View CA 94040

Can Kirmizibayrak is with Google, Inc., Mountain View CA 94043

Jeffrey Schlosser and Dimitre Hristov have financial interests in SoniTrack Systems, Inc.

applications such as radiation therapy and stereotactic surgery where motion management with respect to computerized anatomy models is critical (Comeau et al., 2000, Lange et al., 2004, Schlosser et al., 2010, Schlosser et al., 2012), apparent displacements caused by transducer motion must be decoupled from actual underlying anatomy motion by transforming images to a "world" reference frame—a frame that does not change with probe motion. This transformation requires tracking of the probe's world position using a position sensor, as well as knowledge of the spatial relationship between the sensor and the US image. The process of identifying the unknown relationship between sensor and image is referred to as US spatial calibration.

Previous methods for spatial calibration have mostly relied on scanning customized calibration phantoms containing a plurality of structures with known geometry (Mercier et al., 2005, Hsu et al., 2007). Examples of such structures include wire crosses (Detmer et al., 1994, Périá et al., 1995), 2D shapes (Sato Y Fau - Nakamoto et al.), and Z-wires (Lindseth et al., 2003). The spatial positions of the structures define a set of known coordinates in the world reference frame. Positions of the same structures within the US images are found by extracting a set of common features via image segmentation. A least-squares minimization is then used to find the parameters of the spatial calibration matrix that minimize the distance between the known world coordinates and the image-based coordinates transformed into the world frame.

It is possible to perform spatial calibration without prior knowledge of the spatial relationship between phantom structures. One such approach involves imaging the wall of a phantom and segmenting US image points along the wall (Prager et al., 1998, Rousseau et al., 2005, Hartov et al., 2010). The spatial calibration matrix for which the US points are transformed most closely to a single plane in the world frame is selected. In another approach, a magnetic resonance imaging (MRI) scan of the phantom is obtained along with a set of US images (Blackall et al., 2000). The spatial calibration matrix is chosen such that similarity between the MRI and US scans is maximized.

The calibration approaches mentioned thus far are applicable to both 2D (planar) and 3D (volumetric) US imaging systems. By focusing solely on the calibration of 3D images, a new approach to spatial calibration is made possible. In this approach, a mathematical system of transformation loops (known as "hand eye calibration") is formed by relating a set of 3D US volumes acquired from several tracked probe positions (refer to Section III. B for details). Lange et al. and Bergmeir et al. present calibrations using the hand eye framework, relating sets of US volumes using segmentation of specific image features (Lange and Eulenstein, 2002, Bergmeir et al., 2009). Huang et al. relate US volumes using 3D intramodality image registration (Huang et al., 2010), an approach valid for arbitrary calibration phantoms.

Unlike other spatial calibration methods, hand eye calibration using intramodality image registration does not rely on high-precision custom-designed calibration phantoms (which can be expensive), does not require knowledge or localization of physical phantom points (which can be error-prone and sensitive to phantom manufacturing practices as well as aging of the phantom material), does not require image segmentation (which can be unreliable due to US speckle and image noise), and does not require inter-modality image registration (which can be inaccurate due to differing physical contrast mechanisms between imaging modalities). However, the following important questions regarding the practical implementation and use of US hand eye calibration methods remain unaddressed: (1) Which similarity metric should be used to register US volumes? (2) How can the accuracy of the calibration be validated without using a custom-designed calibration phantom? (3) After how many collected images has the calibration achieved adequate convergence? In other

words, what is the stopping criteria for collecting image poses? (4) How can voxel coordinate changes caused by 3D region-of-interest (ROI) imaging—offered by modern matrix array transducers—be taken into account without re-calibration? (5) How do calibrations performed with different phantoms compare? The purpose of this paper is to address each of these questions, thus providing users a set of practical guidelines for automatically spatially calibrating 3D US systems using an arbitrary phantom.

III. Method and Materials

III. A. Definition of reference frames

A voxel coordinate in a 3D US ROI volume is transformed into a physical point in the relevant world reference frame (Figure 1) using the following equation:

$${}^w p = {}^w T^{tr} * {}^{tr} T^{sn} * {}^{sn} T^{im} * {}^{im} A^{roi} * {}^{roi} p, \quad (1)$$

where ${}^a p$ is a data point in frame a , ${}^a T^b \in \mathbb{R}^{4 \times 4}$ is a homogeneous rigid transformation (rotation and translation) from frame b to frame a , ${}^a A^b \in \mathbb{R}^{4 \times 4}$ is an affine transformation between frame a and b (rotation, translation, and scaling), and the frames w , tr , sn , and im , and roi correspond to the world, tracking device, sensor, US image (volume), and US ROI, respectively. In image guided therapy, the relevant "world" reference frame is either the frame of the computerized patient model or the delivery device.

The transformation between voxel coordinates in the ROI and physical image coordinates is the subject of Section III. C (${}^{im} A^{roi}$). The transformation ${}^{tr} T^{sn}$ is provided by an optical, magnetic, or mechanical tracking device that monitors the position of a sensor rigidly fixed to the US probe. An unknown transformation, ${}^w T^{tr}$ relates the world and tracker frames, which can be solved easily using well-established techniques (Horn et al., 1988). US spatial calibration is used to find the final unknown transformation ${}^{sn} T^{im}$. Note that spatial calibration requires the use of any stationary reference frame, and since the tracker frame (tr) is stationary and can be used for this purpose, finding the world transformation ${}^w T^{tr}$ is not necessary to perform spatial calibration.

III. B. Hand eye US calibration theory

Figure 2 shows a series of n US volumes collected from different tracked probe positions. A unique advantage of 3D US (compared with 2D US) is that an explicit transformation between arbitrary US volume i and arbitrary US volume j can be found using intramodality image registration (Section III. E). The transformation between volumes i and j can also be written using the following loop equation (see Figure 2):

$${}^{im(i)} T^{im(j)} = {}^{im(i)} T^{sn(i)} * {}^{sn(i)} T^{tr} * {}^{tr} T^{sn(j)} * {}^{sn(j)} T^{im(j)}, \quad (2)$$

where ${}^{im(i)} T^{sn(i)} = {}^{im(j)} T^{sn(j)} = {}^{im} T^{sn}$, since the tracking sensor is fixed with respect to the US imaging probe. Right-multiplying Eq (2) by $({}^{sn} T^{im}) = {}^{im} T^{sn}$, we arrive at

$${}^{im(i)} T^{im(j)} * {}^{im} T^{sn} = {}^{im} T^{sn} * {}^{sn(i)} T^{tr} * {}^{tr} T^{sn(j)}. \quad (3)$$

Equation (3) has the form

$$A^{ij} X = X B^{ij}, \quad (4)$$

where $A^{ij} = {}^{im(i)} T^{im(j)}$, $B^{ij} = {}^{sn(i)} T^{tr} * {}^{tr} T^{sn(j)}$, $X = {}^{im} T^{sn}$, and $(i, j) \in \{1, 2, \dots, n\}$. The system represented in Eq (4) is commonly encountered in the field of robotics and is known as "hand-eye calibration" (Horaud and Dornaika, 1995, Daniilidis, 1999). To solve the system,

we use a unit quaternion to compute the rotation component in closed form, then solve for translation using linear least-squares (Horaud and Dornaika, 1995).

III. C. Relating ROI voxels to image coordinates

In 3D US imaging there is a tradeoff between spatial resolution, ROI size, and frame rate. For image guidance applications, the smallest ROI should be chosen that encompasses the therapy target, relevant healthy tissues and/or surgical tools, and corresponding motion margins in order to maximize the frame rate for a given spatial resolution. However, adjusting the 3D ROI and the spatial resolution changes the spatial relationship between the voxel coordinates within the field of view and the sensor mounted to the probe. Instead of re-calibrating the probe each time the ROI or spatial resolution is adjusted, we calibrate the probe with respect to a fixed "image reference frame." The image reference frame is related to voxel coordinates within the ROI based on the spatial imaging resolution in each direction as well as the location and size of the ROI.

For each US image, the position and orientation of the image reference frame with respect to the ROI frame are provided by the US machine as translations, rotations, and scaling factors, combining to form the affine transformation ${}^{im}A^{roi}$ (Equation 1). Note that the ROI frame has voxel units, while the image reference frame has physical coordinates (e.g. millimeters); hence, the need for scaling factors. Figure 3 (a) illustrates the image reference frame and the ROI frame in the 2D case. Figure 3 (b) depicts the ROI voxels superimposed on the full US image volume using the transformation ${}^{im}A^{roi}$.

III. D. Experimental equipment, data acquisition, and implementation

All experiments described in the sections below were performed using two phantoms: (1) CIRS Multi-modality pelvic phantom Model 048 with 3 gold fiducial markers embedded in the phantom prostate; (2) calibration phantom for the Elekta Clarity radiotherapy positioning system (Elekta AB, Stockholm, Sweden), containing a series of well-defined cylinders and a central sphere. The US machine was a Philips iU22 with X6-1 matrix transducer. Image volume size was $432 \times 326 \times 183$ with voxel dimensions $0.70 \text{ mm} \times 0.63 \text{ mm} \times 1.07 \text{ mm}$. The tracking system was an Atracsys EasyTrack 500 3-camera system with active infrared light emitting diode sensors. Manufacturer specified tracker error was $<0.3 \text{ mm}$ within the working volume.

Rigid volumetric image registrations were performed using the MERIT toolbox (Boehler et al. 2011) within the MevisLab software framework (Heckel et al. 2009). US volumes were streamed in real-time using the iU22's Digital Navigation Link and synchronized with optical tracking data using customized MevisLab modules written in C++. The user interface for calibration was implemented using MevisLab's custom MDL language with additional scripting using Python.

III. E. Image registration

US image registration must be robust with respect to varying image appearance due to probe position and angle. Five similarity metrics for rigid image registration were evaluated: normalized mutual information (NMI), normalized cross correlation (NCC), localized cross correlation (LCC), normalized gradient field (NGF), and sum of squared differences (SSD) (Boehler et al. 2011). To compare the accuracy of different similarity metrics, we used the approach outlined in Figure 4. The premise of the approach is that A^{ij} (image transformation) and B^{ij} (sensor transformation) in equation (4) measure the same underlying displacement, but in different reference frames (Chen, 1991). The manufacturer-reported error for measuring sensor poses using the optical tracker was relatively low ($<0.3 \text{ mm}$, see III. D), so the sensor transformations B^{ij} are used as ground truth. The magnitudes of the

rotation and translation for motion A^{ij} are compared to the ground truth displacements in B^{ij} to infer the image error magnitudes (see Appendix VIII. A for details). Registrations were performed with an initial guess that, on average, fell within approximately 2 mm of the final registration transformation achieved. Forty images were collected for each phantom using a variety of transducer angles and positions, producing 39 image registrations per phantom (the first image was registered to all following images).

Comparison of transformations A^{ij} and B^{ij} can also be used as a criteria to reject bad data during the calibration process (Huang et al., 2010). If the rotation or translation magnitudes between A^{ij} and B^{ij} differ beyond a specified threshold, a bad image registration or sensor reading is suspected and that data is discarded. Potential causes of bad data could include image artifacts due to inadvertent probe motion during image acquisition (poor registration), insufficient overlap between US volume pairs (poor registration), or in the case of magnetic tracking, presence of metal or radiofrequency interference within the sensor field (poor sensor reading). The thresholds used for data rejection were 1.5 mm and 1.5 degrees. We found that for a given number of collected images, those thresholds retained satisfactory data quality while still maintaining enough accepted image/sensor pairs to produce good calibrations (see Section IV. A).

Image registration algorithms generally cannot distinguish regions outside image boundaries as spurious image domains. Such regions affect registration results as the optimization can try to treat ‘black’ (i.e. outside) regions as data points and try to align these regions in addition to the image voxels. To avoid this we implemented an automatic method to find the largest volumetric box contained in overlapping regions of two US volumes using a 3D extension of a 2D solution to the maximal rectangle problem (Vandevoorde, 1998). Registration is only performed using the voxels in the overlapping box.

III. F. Stopping criteria and self-consistency error

To automate the calibration process, a metric for stopping the collection of images must be implemented. We used a self-consistency metric that measured the residual error of each individual image using the “leave one out” method outlined below.

After N images are collected:

For i from 2 to N :

- Find $^{sn}T^{im}$ leaving out image i
- Compute estimate of $^{im(1)}T^{im(i)}$ using $^{sn}T^{im}$.

$$^{im(1)}T_{est}^{im(i)} = ^{im}T^{sn} * ^{sn(1)}T^{tr} * ^{tr}T^{sn(i)} * ^{sn}T^{im}$$

- Compare with image registration $^{im(1)}T^{im(i)}$:

$$error(i) = 1/K \sum_{j=1}^K \left\| \left(^{im(1)}T_{est}^{im(i)} - ^{im(1)}T^{im(i)} \right) * p_j^{im} \right\|$$

$$\text{Average Error} = 1/(N-1) \sum_{i=2}^N error(i)$$

The variable p_j^{im} represents the coordinates of an arbitrary point within the image. We use a set of 125 points ($K=125$) spaced evenly on a grid centered at the center of the image and spanning half the image volume. The error was plotted as a function of the number of images collected. A total of 15 runs were performed with different random permutations of image order.

III. G. Calibration validation

Traditionally, the spatial calibration matrix is validated by imaging a phantom with known points in the world frame. Points are identified in the US image, transformed to the world frame via the spatial calibration matrix, and their locations are compared to the known locations within the phantom. One of the major advantages of using the hand-eye US calibration method is that a customized phantom is not required for the calibration. Therefore it is desired to have a validation method that also does not rely on a custom phantom.

III. G. 1. Method and experiments—We use a new validation method that is based on measurements of relative displacements instead of specific world coordinates (Figure 5). First, a set of US images from various poses is taken of an arbitrary phantom (the phantom used for calibration or a different one). The phantom is then physically displaced to a new location, and a second set of US images is acquired. A reference sensor is placed on the phantom, and the tracker measurement of the reference sensor displacement is used as ground truth. Displacements between spatially localized US images collected in each phantom position (images transformed to the tracker reference frame) are measured and compared to ground truth displacements.

Image-based displacements (left box, Figure 5) were measured between 5 images collected in phantom position 1 and 5 images collected in phantom position 2 (total of 25 validation measurements per phantom). The image displacements were measured in two ways:

1. Manually selecting image points. To perform validation using this approach, three users were asked to select the locations of 3 fiducial markers within each of the 10 US volumes of the pelvic phantom. Distances were measured between marker locations in each phantom position transformed to the tracker reference frame.
2. Automatic image registration with manual rejection of outliers using volumetric overlay. Volumes in each phantom position were registered automatically using LCC, the best-performing similarity metric (see results in Section IV. A). After registration, 3 users were asked to rank the quality of the registrations on a scale of 1–3 using a volumetric overlay. Rank 1 corresponded to “aligned” volumes, rank 2 was “perhaps misaligned”, and rank 3 was “definitely misaligned.” Images ranked 2 or 3 were excluded from the validation data set. For a fair comparison with manual point picking, distances were measured between three “pseudo” marker locations in the US volumes—points corresponding with the locations of the manually selected markers above. Pseudo points in one US volume were displaced according to the image registration transformations, and distances were measured between the original points and the displaced points.

III. G. 2. Displacement measurement consistency—Consistency of each image displacement measurement method was investigated in a separate set of experiments. For manual point picking, 10 individual users were asked to select the locations of 3 fiducial markers within a single US volume, and the inter-user spread was evaluated. For volumetric overlay, 50 image volumes were displaced a random amount from their original positions in 6 degrees of freedom and individually overlaid on the original image. Three users ranked the

resulting overlays on the 1–3 scale described above for each phantom, and the inter- and intra-user spread was evaluated. For the volumetric rankings, ground truth physical displacement of three “pseudo” marker locations was measured based on the randomly generated displacements that were ranked “1” by each user (see #2 above).

IV. Results

IV. A. Image registration

Figure 6 shows a comparison of the accuracy of 5 similarity metrics for rigid image registration of US volumes collected from phantoms.

Table 1 statistically compares the similarity metrics in Figure 6 using Wilcoxon rank-sum tests.

When considering both translation and rotation registration errors across both phantoms, LCC and NMI statistically outperform NCC, SSD, and NGF. LCC produced slightly lower mean errors than NMI in both translation and rotation, thus LCC was used as the similarity metric of choice for subsequent experiments. The 1.5 mm and 1.5 degree motion rejection thresholds (Section III. E) discarded 4/39 prostate phantom LCC registrations and 3/39 Clarity phantom LCC registrations in the hand-eye calibration process.

IV. B. Stopping criteria self consistency error

Figure 7 plots the stopping criteria error vs. the number of images collected for each phantom, using LCC as the similarity metric. To achieve 1.5 mm error within the calibration data set, on average only 6 images were required for both the pelvic and Clarity phantoms. To achieve 1.25 mm mean error within the calibration data set, on average 9 pelvic images and 8 Clarity images were required. The final error was 1.11 mm and 0.94 mm for the pelvic phantom (using 35 non-rejected registrations) and Clarity phantom (using 36 non-rejected registrations), respectively.

IV. C. Calibration validation

Figure 8 depicts the consistency of marker picking (inter-user) and volumetric alignment (combined inter- and intra-user) for measuring image displacements (Section III.G.2). The spread of the data from the marker centroid is shown for each of the three markers. Marker locations in X, Y, and Z directions are all contained on the vertical axis of the figure.

Figure 9 shows the ground truth error for volumes ranked 1, 2, and 3 by all three users (Section III.G.2). Root mean square (RMS) ground truth errors for volumes ranked ‘1’ were 0.7 mm and 0.5 mm for the pelvic and Clarity phantom, respectively. A Wilcoxon rank-sum test revealed a statistically significant difference ($p=0.003$) between pelvic and Clarity phantom perturbations on volumes ranked ‘1’, meaning that users could more accurately judge volumetric alignment on the Clarity phantom—with its clearly defined, high-contrast structures—than the pelvic phantom.

Figure 10 compares the different validation methods across different phantoms and different users (Section III.G.I). For volumetric validation data labeled ‘all users,’ a consensus-based approach was used where the volumetric alignment was included if at least 2/3 users ranked the volume ‘1’. For point-based validation data labeled ‘all users,’ the centroid among all 3 users was used, and data was excluded for image pairs not included in the volumetric consensus. The volumetric consensus among users excluded 12/25 and 6/25 volume pairs for the pelvic and Clarity phantom, respectively.

Validating the pelvic phantom calibration across all users, RMS ground truth error was 2.0 mm, 1.7 mm, and 1.3 mm when validating with pelvic point selection (PPS), pelvic volumetric alignment (PVA), and Clarity volumetric alignment (CVA), respectively. Validating the Clarity phantom calibration across all users, RMS ground truth error was 2.4 mm, 2.0 mm, and 1.4 mm when validating with PPS, PVA, and CVA, respectively. Table 2 compares the individual validation methods in Figure 10 using a set of Wilcoxon rank-sum tests. A separate Wilcoxon rank-sum test revealed no statistically significant difference ($p=0.79$) between calibrations performed with the pelvic and Clarity phantoms when validated across all users with CVA, the most accurate validation metric.

V. Discussion

We presented a new implementation of spatial US calibration using 3D intramodality image registration. The unique combination of: (1) using motion-based data rejection criteria, (2) performing registration only over overlapping image regions, and (3) computing the “self-consistency error” online after every collected image, enables the calibration to proceed fully automatically. The user’s only role is to collect US volumes (and corresponding sensor tracking information) until an acceptable “self-consistency” error threshold is met. The calibrations validated in Figure 10 use all non-rejected registrations (35 for the pelvic phantom and 36 for the Clarity phantom), since more images/registrations generally yield more accurate calibrations (see Figure 7). In our implementation, typical time for image collection, automatic registration, and self-consistency check for each calibration image ranged from 10–15 seconds (using NMI) up to one minute (using LCC). Note that image registration code was not optimized for computational speed.

When performing 3D image registration during calibration, one may wonder why the choice of similarity metric matters since poor registrations are discarded using the automatic rejection criteria (Section III. E). The reason is that motion-based rejection only accounts for two out of six degrees of freedom (DOFs) in the image registrations, since it is only possible to discard data based on the *magnitude* of the translation and the rotation (see Appendix VIII. A). Images may pass the 2-DOF rejection criteria (translation and rotation magnitude), but the other 4 DOFs of the registration could still be incorrect. More accurate similarity metrics minimize the chances of a mis-registration with respect to the remaining 4 DOF. Although the image registration results in Figure 6 only capture the registration accuracy in 2 DOF, the 39 registrations per metric were performed from a wide range of angles, thus varying the particular 2 DOF assessed (Figure 4) and providing a comprehensive picture of the overall registration accuracy of each metric. For the above reasons, and based on the results of Figure 6 and Table 1, it is recommended that LCC or NMI is used for 3D intramodality registration of US phantom images. The NMI metric is commonly used for registering multi-modality images because of its robustness to image appearance. Our results support the intuition that the robustness of NMI (and LCC, which has related properties to NMI) contributes to superior results for registration of 3D US images compared to other methods, due to the varying appearance of US features when imaged from different transducer angles.

In selecting an approach to measure image-based displacements for use with the proposed validation method (Figure 5), it is important to note that unlike calibration, image registration with automatic rejection of outliers is not a valid scheme. Since the tracker-measured displacement is used as ground truth in the proposed validation method (Figure 4), using the tracker-measured phantom displacement as part of an automatic rejection criterion could bias the data. Instead, two approaches involving manual intervention were proposed: point picking and ranking of volumetric alignment. Figure 8 shows that the data spread for point-picking is highly dependent on the position of the particular point (marker) within the

US volume, but in any case, volumetric alignment ranking on either phantom is more consistent than selecting individual points within US images.

The calibration validation results in Figure 10 demonstrate that across all users, volumetric alignment ranking is not only more consistent, but also more accurate in determining physical phantom displacements than point selection. Phantom displacement measurement accuracy is affected by the quality of the calibration matrix (the object of the validation), the accuracy of the tracker, and the uncertainty associated with measuring image displacements (Figure 5). For comparative reference, optical tracker error was <0.3 mm (Section III. D), RMS ground truth volumetric alignment error was 0.5 mm for the Clarity phantom (Figure 9), and RMS ground truth displacement measurement error was <1.5 mm as evaluated by CVA (Figure 10). We focus on CVA results because CVA was the most consistent (Figure 8) and accurate (Figure 9) validation metric tested. Our error compares favorably to other implementations of US hand eye calibration: Lange et al. report 3.3 mm RMS error (Lange and Eulenstein, 2002) using 5 images, Huang et al. report 2.4 mm RMS error using 56 images (Huang et al., 2010), and Bergmeir et al. report 3.5 mm mean error using 12 images (Bergmeir et al., 2009). The significant improvement (<1.5 mm RMS error using ~ 36 images) relative to previously reported errors may be attributed to our motion-based data rejection criteria, use of LCC for image registrations (instead of NCC as in (Huang et al., 2010)), high-end Philips X6-1 imaging transducer, and/or our accurate and reproducible validation method (CVA) based on manual rankings of volumetric image alignment.

Does the particular choice of US phantom matter when performing the hand-eye calibration or when validating the calibration? When ranking volumes for validation, users are able to rank volumetric alignment of the Clarity phantom with significantly lower ($p=0.003$) ground truth error than alignment with the pelvic phantom (Figure 9). Accordingly, Table 2 indicates that validations performed using CVA have significantly lower errors than validations using PVA. Calibrations using the Clarity and pelvic phantoms do not result in significantly different validation errors as evaluated by CVA (Figure 10). We conclude that for manually evaluating image displacements using volumetric alignment (for calibration validation), phantoms with clearly defined structures such as the Clarity yield lower errors than generic US phantoms. For calibration, with the help of automatic motion-based data rejection, different phantoms produce similar results, although some phantoms may converge to a lower error with less images than others (Figure 7).

VI. Conclusion

We evaluated a new automatic technique for spatial calibration of 3D US images using intramodality image registration. In the evaluation process, we found that NMI and LCC yielded the best 3D US image registrations, and that manual alignment of US volumes is more reliable than manually selecting points within the volumes for calibration validation. RMS spatial calibration error was <1.5 mm when evaluated using a novel validation technique based on phantom displacements and volumetric image alignment, a significant improvement relative to previously reported results. Comparison of two different phantoms for calibration and for validation revealed significant differences for validation ($p=0.003$) but not for calibration ($p=0.795$).

Acknowledgments

The authors acknowledge the Stanford BioX Interdisciplinary Initiatives Program, Philips Ultrasound, and NIH grant IR41CA174089 for providing support of this research. Vijay Shamdasani and Stephen Metz are employees of Philips Ultrasound.

Appendix

VIII. A. Rotation and translation from rigid transformation matrices

Rotation and translation magnitudes are extracted from the rigid transformations A^{ij} and B^{ij} using a screw motion convention (Bottema and Roth, 1979, Chen, 1991). Let R and t denote the 3×3 rotation matrix and 3×1 translation matrix associated with a rigid body motion. The rotation angle ϕ and unit vector rotation axis r corresponding to rotation matrix R are given below

$$\phi = \cos^{-1} \left(\frac{\text{trace}(R) - 1}{2} \right) \quad r = \frac{1}{2\sin(\phi)} \begin{bmatrix} R_{3,2} - R_{2,3} \\ R_{1,3} - R_{3,1} \\ R_{2,1} - R_{1,2} \end{bmatrix}$$

where $R_{i,j}$ indicates element (i,j) in matrix R . Rotation magnitude θ and translation magnitude d are given by

$$\theta = |\phi| \quad d = |t \cdot r|$$

References

- Bergmeir C, Seitel M, Frank C, Simone RD, Meinzer H-P, Wolf I. Comparing calibration approaches for 3D ultrasound probes. *International journal of computer assisted radiology and surgery*. 2009; 4:203–213. [PubMed: 20033620]
- Blackall, J.; Rueckert, D.; Maurer, C.; Penney, G.; Hill, D.; Hawkes, D. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2000*. Springer; 2000. An image registration approach to automated calibration for freehand 3D ultrasound.
- Boehler T, van Straaten D, Wirtz S, Peitgen H-O. A robust and extendible framework for medical image registration focused on rapid clinical application deployment. *Computers in biology and medicine*. 2011; 41:340–349. [PubMed: 21507391]
- Bottema, O.; Roth, B. *Theoretical kinematics*. Dover Publications; 1979.
- Chen, HH. *Proceedings CVPR'91*, IEEE Computer Society Conference on. IEEE; 1991. A screw motion approach to uniqueness analysis of head-eye geometry. *Computer Vision and Pattern Recognition*, 1991; p. 145-151.
- Comeau RM, Sadikot AF, Fenster A, Peters TM. Intraoperative ultrasound for guidance and tissue shift correction in image-guided neurosurgery. *Medical Physics*. 2000; 27:787–800. [PubMed: 10798702]
- Daniilidis K. Hand-Eye Calibration Using Dual Quaternions. *The International Journal of Robotics Research*. 1999; 18:286–298.
- Detmer PR, Bashein G, Hodges T, Beach KW, Filer EP, Burns DH, Strandness JR, D E. 3d ultrasonic image feature localization based on magnetic scanhead tracking: In vitro calibration and validation. *Ultrasound in medicine & biology*. 1994; 20:923–936. [PubMed: 7886852]
- Hartov A, Paulsen K, Ji S, Fontaine K, Furon M-L, Borsic A, Roberts D. Adaptive spatial calibration of a 3D ultrasound system. *Medical Physics*. 2010; 37:2121–2130. [PubMed: 20527545]
- Heckel F, Schwier M, Peitgen HO. Object oriented application development with MeVisLab and Python. *Lecture Notes in Informatics*. 2009; 154:1338–1351. (Informatik: Im Focus das Leben).
- Horaud R, Dornaika F. Hand-Eye Calibration. *The International Journal of Robotics Research*. 1995; 14:195–210.
- Horn BKP, Hilden HM, Negahdaripour S. Closed-Form Solution of Absolute Orientation Using Orthonormal Matrices. *Journal of the Optical Society of America a-Optics Image Science and Vision*. 1988; 5:1127–1135.
- Hsu P-W, Prager RW, Gee AH, Treece GM. Freehand 3D Ultrasound Calibration: A Review. 2007

- Huang, X.; Gutiérrez, LF.; Stanton, D.; Kim, PC.; Jain, A. Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on. IEEE; 2010. Image registration based 3D TEE-EM calibration; p. 1209-1212.
- Lange T, Eulenstein S. Calibration of swept-volume 3-D ultrasound. *MIUA Proc.* 2002; 99:29–32.
- Lange, T.; Eulenstein, S.; Hünerbein, M.; Lamecker, H.; Schlag, P-M. Augmenting Intraoperative 3D Ultrasound with Preoperative Models for Navigation in Liver Surgery *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*. In: Barillot, C.; Haynor, D.; Hellier, P., editors. Springer Berlin / Heidelberg; 2004.
- Lindseth F, Tangen GA, Langø T, Bang J. Probe calibration for freehand 3-D Ultrasound. *Ultrasound in medicine & biology*. 2003; 29:1607–1623. [PubMed: 14654156]
- Mercier L, Langø T, Lindseth F, Collins LD. A review of calibration techniques for freehand 3-D ultrasound systems. *Ultrasound in Medicine & Biology*. 2005; 31:143–165. [PubMed: 15708453]
- Netsch T, Roesch P, Muiswinkel A, Weese J. Towards Real-Time Multi-Modality 3-D Medical Image Registration. *International Conference on Computer Vision (ICCV'01)*. 2001; 1:718–725.
- Péria, O.; Chevalier, L.; François-Joubert, A.; Caravel, J-P.; Dalsoglio, S.; Lavallée, S.; Cinquin, P. Using a 3D position sensor for registration of SPECT and US images of the kidney. In: Ayache, N., editor. *Computer Vision, Virtual Reality and Robotics in Medicine*. Springer Berlin / Heidelberg; 1995.
- Prager RW, Rohling RN, Gee AH, Berman L. Rapid calibration for 3-D freehand ultrasound. *Ultrasound in medicine & biology*. 1998; 24:855–869. [PubMed: 9740387]
- Rousseau F, Hellier P, Barillot C. Confhustus: A robust and fully automatic calibration method for 3D freehand ultrasound. *Medical Image Analysis*. 2005; 9:25–38. [PubMed: 15581810]
- Sato Y Fau - Nakamoto M, Nakamoto M Fau - Tamaki Y, Tamaki Y Fau - Sasama T, Sasama T Fau - Sakita I, Sakita I Fau - Nakajima Y, Nakajima Y Fau - Monden M, Monden M Fau - Tamura S, Tamura S. Image guidance of breast cancer surgery using 3-D ultrasound images and augmented reality visualization.
- Schlosser J, Salisbury K, Hristov D. Telerobotic system concept for realtime soft-tissue imaging during radiotherapy beam delivery. *Medical Physics*. 2010; 37:6357–6367. [PubMed: 21302793]
- Schlosser J, Salisbury K, Hristov D. Online Image-based Monitoring of Soft-tissue Displacements for Radiation Therapy of the Prostate. *International journal of radiation oncology, biology, physics*. 2012
- Vandevoorde D. The maximal rectangle problem. *Dr. Dobb's journal*. 1998; 23

Figure 1. Transformation chain relating voxel coordinates in the ROI to physical coordinates in the world frame.

Figure 2.

Hand eye US calibration. The transformation between arbitrary US volume i and j can be found indirectly via the tracked sensor (probe) positions, or directly via 3D intramodality image registration.

Figure 3. Relationship between ROI and image reference frames. (a) Schematic. (b) ROI volume (red) superimposed on full US volume.

Figure 4.

Method for comparing image registration errors. Motions measured by the tracker are used as ground truth data to which the image registrations are compared.

Figure 5.

Method for validating US spatial calibration accuracy using measurements of relative displacements. Movements measured by the reference sensor are compared to displacements measured in images that are transformed to the tracker frame using the spatial calibration matrix.

Figure 6.

Comparison of rigid US registration similarity metrics across calibration phantoms. Combined data includes data from both phantoms. Horizontal dotted red lines indicate thresholds for rejecting registrations based on the motion rejection criteria (Section III. E). In this figure and all following figures containing box plots the central mark is the median, the box edges are 25th and 75th percentiles, whiskers extend to most extreme data points not considered outliers, and outliers are plotted individually. Points are drawn as outliers if they are larger than $q3 + 1.5 \cdot (q3 - q1)$ or smaller than $q1 - 1.5 \cdot (q3 - q1)$, where $q1$ and $q3$ are the 25th and 75th percentiles, respectively.

Figure 7. Self-consistency error as a function of number of images collected for each phantom. Thin blue lines show results with particular random permutations of image order, and thick red lines show the mean across all permutations.

Figure 8.

Data spread for point selection and for volumetric alignment (Section III.G.2). Spreads indicate how consistently users can match images using point selection and using volumetric alignment for purposes of calibration validation. Markers in parenthesis are “pseudo” markers that are used for purposes of comparing volumetric alignment with point selection.

Figure 9.
Ground truth errors for randomly misaligned US volumes according to user ranking and phantom (Section III.G.2).

Figure 10.

Comparison of point-based and volumetric-based validation methods across users and phantoms. Validations for a calibration performed with the pelvic phantom are shown on the left, and validations for a Clarity phantom calibration are shown on the right.

Table 1

Wilcoxon rank-sum tests with null hypothesis that the median difference between the combined LCC data (both phantoms) and the other image registration similarity metrics is zero. P-values shown in italics reject the null hypothesis.

	Translation p-value	Rotation p-value
NCC vs LCC	<i>p<0.001</i>	<i>0.044</i>
SSD vs LCC	<i>p<0.001</i>	0.089
NMI vs LCC	0.18	0.76
NGF vs LCC	0.27	<i>0.006</i>

Table 2

Wilcoxon rank-sum tests with null hypothesis that the medians of the data labeled ‘all users’ in Figure 10 are equal. P-values shown in italics reject the null hypothesis. PPS = pelvic point selection, PVA = pelvic volumetric alignment, and CVA = Clarity volumetric alignment.

	PPS vs. PVA p-value	PPS vs. CVA p-value	PVA vs. CVA p-value
Pelvic phantom calibration	0.053	<i>p<0.001</i>	<i>p<0.001</i>
Clarity phantom calibration	0.10	<i>p<0.001</i>	<i>p<0.001</i>